Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions

Lifeng Luo,[1,2] Eric F. Wood,[2] and Ming Pan[2]

[1]   This study uses a Bayesian approach to merge ensemble seasonal climate forecasts generated by multiple climate models for better probabilistic and deterministic forecasting. Within the Bayesian framework, the climatological distribution of the variable of interest serves as the prior, and the likelihood function is developed with a weighted linear regression between the climate model hindcasts and the corresponding observations. The resulting posterior distribution is the merged forecast, which represents our best estimate of the variable, including its mean and variance, given the current model forecast and knowledge about the model's performance. The handling of multimodel climate forecasts and nonnormal distributed variables, such as precipitation, are two important extensions toward the application of the Bayesian merging approach for seasonal hydrological predictions. Two examples are presented as follows: seasonal forecast of sea surface temperature over equatorial Pacific and precipitation forecast over the Ohio River basin. Cross validation of these forecasts shows smaller root mean square error and smaller ranked probability score for the merged forecast as compared with raw forecasts from climate models and the climatological forecast, indicating an improvement in both deterministic and probabilistic forecast skills. Therefore there is great potential to apply this method to seasonal hydrological forecasting.

1.   Introduction

[2]   Seasonal climate predictions using comprehensive coupled ocean-atmosphere-land models are now being made routinely at a number of operational weather and climate centers around the world, such as the National Centers for Environmental Prediction (NCEP) [*Kanamitsu et al.*, 2002], International Research Institute for Climate Prediction (IRI) at Columbia University [*Barnston et al.*, 2003], and the European Centre for Medium-Range Weather Forecast [*Palmer et al.*, 2000, 2004]. This development and capability is primarily attributed to advances in understanding the interaction between the atmosphere, ocean, and land at seasonal-to-interannual timescales [*Koster et al.*, 2000] as well as tremendous increases in computing power. Practically speaking, improvements in seasonal predictability are to a large extent because of a better understanding of the role of sea surface temperature (SST) variability [for example, El Niño, Southern Oscillation (ENSO)] in the climate system and advances in its observation, while recognizing

that land surface conditions do contribute to improvements in predictions over some regions of the world [*Koster et al.*, 2000].

[3]   Seasonal predictions of climate variables, especially precipitation and air temperature, have a great value to society [*Jones et al.*, 2000; *Schneider and Garbrecht*, 2003; *Everinghama et al.*, 2002; *Palmer*, 2002] and are fundamental to the Coordinated Observation and Prediction of the Earth System activity of the World Climate Research Program. Seasonal predictions of precipitation, and in turn predictions of soil moisture and streamflow, can have great values to our society. Agriculture, water resource management, and energy and transportation sectors are a few among many others that will benefit through appropriate planning given useful seasonal predictions. However, the current skill of seasonal hydrological forecasts is still limited and far from meeting the society's needs. A research interest and priority therefore is to understand the predictability of the climate system at seasonal-to-interannual timescales and to improve seasonal forecast skills. The path that leads to increasing forecast skill involves improving the climate model physics, resolution, parameterizations for unresolved processes, etc., which are constantly carried out at climate modeling centers. One area which has received less attention is the development of statistical postprocessing methods to achieve the best possible prediction with the current models [*Coelho et al.*, 2004; *Stephenson et al.*, 2005]. This paper contributes to addressing this need.

[1]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey, USA.
[2]Environmental Engineering and Water Resource, Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.

[4] One major advance in statistical postprocessing techniques is the development of the multimodel super-ensemble concept [*Krishnamurti et al.*, 2000]. Because of the stochastic nature of the climate system, seasonal forecasts would be better expressed in a probabilistic manner. Although one can provide a single-valued (deterministic) forecast, information about uncertainties should always be included in the forecast. The basic method for addressing forecast uncertainties from deterministic dynamic climate models is through analyzing their ensemble members. Various ensemble generation methods have been used in operational numerical weather forecasts and in many climate studies, including the lagged average forecasting method [*Hoffman and Kalnay*, 1983] and the breeding method [*Toth and Kalnay*, 1993]. These methods attempt to incorporate the uncertainties in the initial conditions of the climate system. However, there are more uncertainties in the system than the initial conditions, as there are uncertainties due to model formulations. Numerical representations of the climate system have uncertainties when the partial differential equations are expressed and solved over finite grids. Parameterizations used in the models have uncertainties as they try to simulate processes that cannot be fully resolved. These uncertainties can propagate and affect the solution of the system across the entire spectrum of scales. As a result, each climate model tends to have its own climatology that may not reflect the real climate system, and each model also has a forecast skill that varies geographically and with lead time and season. To include all the uncertainties and represent them in the forecast, *Krishnamurti et al.* [2000] proposed a multimodel super-ensemble approach that utilizes multiple models for ensemble forecasts and pools all the ensemble members to form a "super-ensemble." They found that the estimates using the super-ensemble outperform all model forecasts for multiseasonal, medium-range weather and hurricane forecasts. Since then, the multimodel super-ensemble concept has received increased attention and has been used in many applications. For instance, *Krishnamurti et al.* [2001] applied the multimodel super-ensemble concept on real-time precipitation forecast using Tropical Rainfall Measuring Mission and SSM/I products; *Kumar et al.* [2003] constructed a multimodel super-ensemble for forecasting tropical cyclones over the Pacific Ocean based on the operational forecast data set; *Williford et al.* [2003] studied the Atlantic hurricane forecast for the year 1999 using the super-ensemble method. The super-ensemble concept has also been implemented in operational seasonal forecasts at different research and operational institutes. *Barnston et al.* [2003] reported the progress on multimodel ensembling in seasonal forecasting at IRI, and *Palmer et al.* [2004] summarized the development of the European multimodel ensemble system for seasonal-to-interannual prediction (a.k.a. DEMETER). All these applications have shown the potential of the multimodel approach in improving various forecasts.

[5] While data-based statistical models are developed for long-lead prediction of ENSO [*Berliner et al.*, 2000], various studies have also been carried out to improve statistical techniques for combining the super-ensembles from dynamic models. Among many others, *Rajagopalan et al.* [2002] used a Bayesian method to optimally combine global seasonal precipitation and temperature forecasts in two different seasons, and they found a general improvement in forecast skills over individual models. *Robertson et al.* [2004] made improvements to the Bayesian scheme of Rajagopalan et al. by reducing the dimensionality of the numerical optimization. They achieved increases in cross-validated forecast skill when combining six atmospheric general circulation model seasonal hindcast ensembles with the revised scheme. *Yun et al.* [2003] introduced a technique for improvement of the long-term forecast skill of the multimodel super-ensemble based on singular value decomposition. Recently, *Coelho et al.* [2004] presented a Bayesian approach for making deterministic forecast of ENSO (SST Niño-3.4 index) based on the dynamical climate forecast from the European Union's (EU) DEMETER project and an empirical statistical model forecast. They were able to show that the combined forecast, using their approach, increases the skill score and provides a reliable estimate of the forecast uncertainty.

[6] In early 2004, the authors started developing a Bayesian merging method for seasonal hydrological predictions, with a focus over the eastern U.S. The developed method has similarities to the approach presented in the work of *Coelho et al.* [2004] but developed independently. The next section presents the Bayesian merging methodology. The major differences in the approach in this paper and that of Coelho et al. is the development of the Bayesian posterior distribution, weighted to reflect the individual model skill, from which multimodel super-ensembles can be generated. The second difference is our ability to handle climate variables from nonnormal distributions within the multimodel framework. We apply the Bayesian multimodel forecasting system to two different applications: the prediction of monthly sea surface temperature over the equatorial Pacific and the prediction of monthly precipitation over the Ohio River basin. These are described in sections 3 and 4. The last section provides discussion and conclusions.

2. Methodology

2.1. Bayes' Theorem

[7] Bayes' theorem provides an approach to update the probability distribution of a variable based on information newly available by calculating the conditional distribution of the variable given this new information. The updated (conditional) probability distribution reflects the new level of belief about the variable. For example, the variable of interest is a quantity θ at a future time (for example, SST at a specific location for a given month). Before it is actually observed, θ would be a random variable, and our knowledge on θ is a probability distribution, i.e., its probability density function (PDF) $p(\theta)$. Without the help of any climate forecast models, $p(\theta)$ would simply be the climatological distribution of θ from historical records. Now with a climate model, a forecast of this variable can be made, and we denote this forecast as y. Then given this new information y, the conditional distribution $p(\theta|y)$ reflects our new belief. Bayes' theorem computes $p(\theta|y)$ as:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \qquad (1)$$

where $p(\theta)$ and $p(y)$ are the unconditional distributions of θ and y. $p(\theta)$ is also referred to as the "prior" distribution of θ
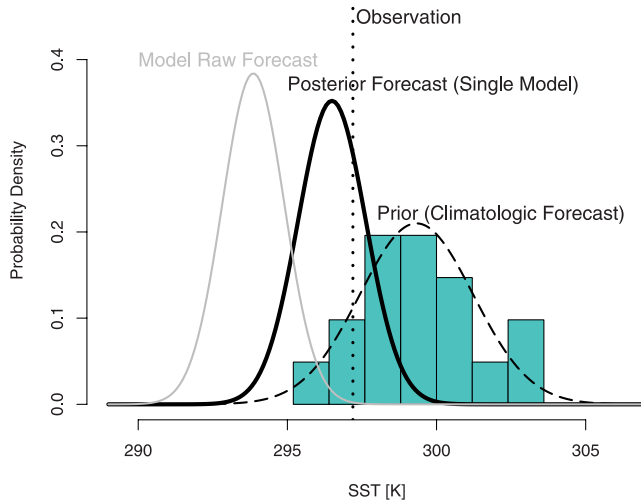
Figure 1. Prior distribution (dashed black) and posterior distribution (solid black) for the forecast of monthly mean SST over selected grid (2.5 × 2.5 centered at 0 and 130°W) of December 1998. The raw forecast (solid gray) distribution estimated from one EU DEMETER climate model is also plotted. The vertical dotted line indicates the actual observation for December 1998. The prior distribution is estimated from data spanning 1982 to 1999 except 1998. The histogram of the 19 months is also shown.

(prior to the new information y), and $p(\theta|y)$ is the "posterior" distribution of θ (posterior to y). $p(y|\theta)$ is referred to as the likelihood function, and it measures how closely y is distributed around θ, i.e., in this example, how skillful the climate model is. The focus of the Bayesian merging method is to develop the probability models for $p(y|\theta)$, $p(y)$, and $p(\theta)$ and to compute the posterior distribution $p(\theta|y)$. The likelihood function is a key step because it measures the discrimination of the model forecasts and thus determines how much information is provided by the model(s). Here the likelihood function is constructed using a linear regression model with normal errors.

[8] To help derive the method and illustrate it with practical applications, we use the following example data set. Let θ be the monthly mean SST forecast variable in a 2.5 × 2.5° box centered at 0° and 130°W for December 1998. This box is within the Niño-3.4 region. Accurate forecasts of SST over the equatorial Pacific are essential for skillful seasonal climate forecasts (*Latif et al.* [1998], *Barnston et al.* [1999], and *Landsea and Knaff* [2000] among others). The dynamical climate model forecasts used here are from the European Union's DEMETER project [*Palmer et al.*, 2004]. Forecasts starting from August for the seven models are used here, such that the lead time for the December forecast is 5 months. The observations of monthly mean SST come from the Reynolds data set [*Reynolds and Smith*, 1994] and have been regridded to match the grid of the DEMETER models.

2.2. Selecting the Prior Distribution

[9] As previously mentioned, an obvious choice for the prior distribution of the sample data set is the climatological distribution of SST from historical observations for that grid during December. Figure 1 shows the histogram of θ for all Decembers from 1982 to 1999, except 1998. A normal

distribution is fit to the data as the prior distribution, $p(\theta) \sim N(\theta_o, \varphi_o)$. θ_o and φ_o are the mean and variance of the distribution, respectively. Other distributions are also possible: for example, in the study of *Coelho et al.* [2004], prediction from an empirical model is used as the prior.

2.3. Modeling the Likelihood Function

[10] The probability model for the likelihood function $p(y|\theta)$ expresses the probability of the forecast y given the observed SST θ and conveys the overall discrimination of the forecast for different SST realizations (observations). The likelihood function can be estimated from the historical performance of the model forecasts based on hindcasts. Figure 2 illustrates climate model forecasts of December SST versus observed SST, and later this relationship is used to construct the likelihood function.

[11] The probability model for the likelihood function can be developed in a number of ways. For this work, because there is a set of model ensembles for a single realized SST, conditional distributions are used to estimate $p(y|\theta)$ using the conditional distribution of the ensemble mean given the observed SST, $p(\bar{y}|\theta)$, and the conditional distribution of the ensembles, given their mean value, $p(y|\bar{y})$, as follows:

$$p(y|\theta) = p(y|\bar{y})p(\bar{y}|\theta) \tag{2}$$

A linear regression model is used to summarize the relationship between the model mean forecast and the observations [*Coelho et al.*, 2003, 2004].

$$\bar{y} = \alpha + \beta\theta + \varepsilon \tag{3}$$

where α and β are the intercept and slope parameters, respectively. The parameters α and β correspond to the bias and scaling error in the model. The variable ε is the residual (zero-mean) of the regression and assumed to be normally distributed, and its variance φ_ε reflects the efficiency of the linear regression. To estimate α and β, a weighted linear regression is used to minimize the variance of ε [*Coelho et al.*, 2003, 2004], and the weights are related to the inverse
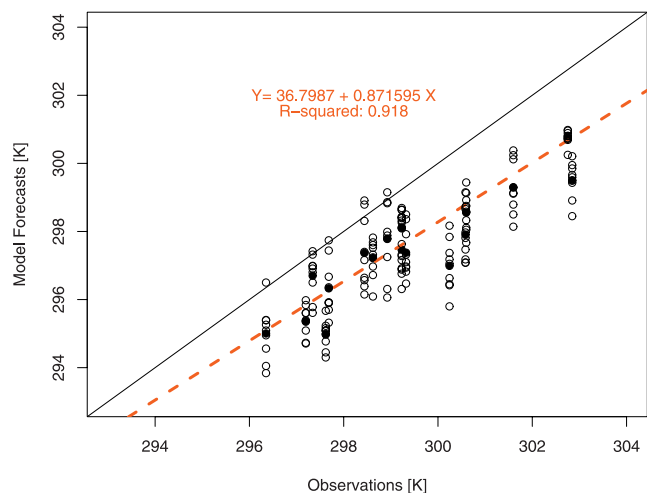

Figure 2. Scatterplot of model forecast of December SST over selected grid versus the observations. The regression line is calculated from the forecast mean with differential weights.

of the ensemble spread. Differential weighting of past forecasts is required because the ensemble spread is not constant throughout the forecasts. A forecast with a larger spread indicates that the mean forecast estimated from the ensemble has a larger uncertainty and therefore should be given a smaller weight.

[12] With the linear model, $\bar{y}$ follows a normal distribution,

$$p(\bar{y}|\theta) \sim N(\alpha + \beta\theta, \varphi_\varepsilon) \tag{4}$$

with mean $\alpha + \beta\theta$ and variance φ_ε the variance of the residuals. Assuming that the ensembles of the current forecast are normally distributed around the mean with variance φ_y, the conditional distribution of y given the ensemble mean can be expressed as:

$$p(y|\bar{y}) \sim N(\bar{y}, \varphi_y) \tag{5}$$

The current forecast variance φ_y is assumed independent of the long-term weighted linear regression error φ_ε, resulting in a likelihood function:

$$p(y|\theta) \sim N(\alpha + \beta\theta, \varphi_y + \varphi_\varepsilon) \tag{6}$$

[13] We note that Coelho et al. [2003, 2004] regressed y against θ directly, which limits the understanding of how the overall variance in likelihood function (and thus the skill from the models) balances the spread of the ensembles for particular forecasts to the efficiency of the forecasts based on the mean of the ensembles. In equation (6), the variance in the likelihood function is composed of two sources of variability, φ_ε represents the efficiency of the linear regression that relates the forecast ensemble mean to the observation, and φ_y is the spread of the ensemble members around the mean. The larger φ_ε is, the less efficient the linear regression in explaining the relation between $\bar{y}$ and θ, which suggests a less skillful climate model resulting in a smaller weight when merged with the prior. The variable φ_y is the variance of the ensembles of the current forecast and needs to represent accurately the uncertainties in the forecast system. There are two factors that contribute to this variance. One is the inherent uncertainty related to natural variability and predictability of the variable. The other factor is related to the probabilistic resolution of the model forecast. This "resolution" is associated with the ability of the model ensemble forecast to separate the forecast distribution from the climatological distribution. For example, if the distribution of the ensemble members is indistinguishable from the climatological distribution, then this ensemble forecast would have poor resolution. Therefore the information provided by this forecast is not very useful statistically. On the other hand, if all or most of the ensembles are clustered, resulting in a very different distribution from the climatological distribution, then the forecast indicates that there is a high probability that the forecast variable will evolve away from its climatology. The resolution of a model forecast depends on many things, including its parameterizations, the state of the climate system, and even how the ensembles are generated. For example, improperly generated ensemble members can be highly correlated and can create a clustering with underestimated uncertainties. All of

these uncertainties are factored into the Bayesian analysis through the magnitude of φ_y as the ensembles are merged with the climatology. The larger φ_y is, the more uncertain the forecast, therefore the less contribution this model provides. This is particularly important when dealing with multiple climate models. Models with small φ_y (small ensemble spread) but low skill will have a large φ_ε, so the variance of $p(y|\theta)$ in equation (6) will still be large.

2.4. Posterior Distribution From Single Model Ensembles

[14] From Bayes' theorem, the posterior distribution $p(\theta|y)$ can be computed and also follows a normal distribution when the prior distribution and the likelihood function are both normal distributions [Lee, 1997]. The posterior distribution is given by:

$$p(\theta|y) = p(\theta|\bar{y}, \varphi_y) \sim N(\theta_p, \varphi_p) \tag{7}$$

with mean θ_p and variance φ_p computed using

$$\begin{aligned} \frac{1}{\varphi_p} &= \frac{1}{\varphi_o} + \frac{1}{\varphi_1} = \frac{1}{\varphi_o} + \frac{\beta^2}{\varphi_y + \varphi_\varepsilon} \\ \frac{\theta_p}{\varphi_p} &= \frac{\theta_o}{\varphi_o} + \frac{\theta_1}{\varphi_1} = \frac{\theta_o}{\varphi_o} + \frac{\beta^2}{\varphi_y + \varphi_\varepsilon}\left(\frac{\bar{y} - \alpha}{\beta}\right) \end{aligned} \tag{8}$$

Note that the posterior distribution is conditioned on the entire distribution of y, not just the mean $\bar{y}$.

[15] Figure 1 shows the computed posterior distribution of the sample data set using equation (8). For comparison, an ensemble forecast with a normal distribution fit to the nine ensemble members is shown as well. The observed SST for the target month is plotted as the vertical dashed line. In this case, the posterior distribution is clearly closer to the observation than either the climatological distribution (the prior distribution) or the raw forecast. Furthermore, an examination of equation (8) shows that the posterior distribution has a smaller variance than the prior φ_o, which indicates that the posterior uncertainty for θ is reduced when prior information is updated with model ensemble forecasts.

2.5. Posterior Distribution From Multimodel Ensembles

[16] If forecasts from multiple climate models are available, with y_i denoting the forecast from model i, then the likelihood function for the forecast from model i can then be expressed in exactly the same manner as equation (6). If these climate models give forecasts with independent errors (φ_ε and φ_y) and with the normality assumption unchanged for the likelihood functions for all models, the posterior distribution is normally distributed as:

$$p(\theta|y_1, y_2, \ldots, y_m) \sim N\left(\theta_{mp}, \varphi_{mp}\right) \tag{9}$$

with mean θ_{mp} and variance φ_{mp} computed using

$$\begin{aligned} \frac{1}{\varphi_{mp}} &= \frac{1}{\varphi_o} + \sum_{j=1}^{m} \frac{\beta_j^2}{\varphi_{y_j} + \varphi_{\varepsilon_j}} \\ \frac{\theta_{mp}}{\varphi_{mp}} &= \frac{\theta_o}{\varphi_o} + \sum_{j=1}^{m} \frac{\beta_j^2}{\varphi_{y_j} + \varphi_{\varepsilon_j}}\left(\frac{\bar{y}_j - \alpha_j}{\beta_j}\right) \end{aligned} \tag{10}$$
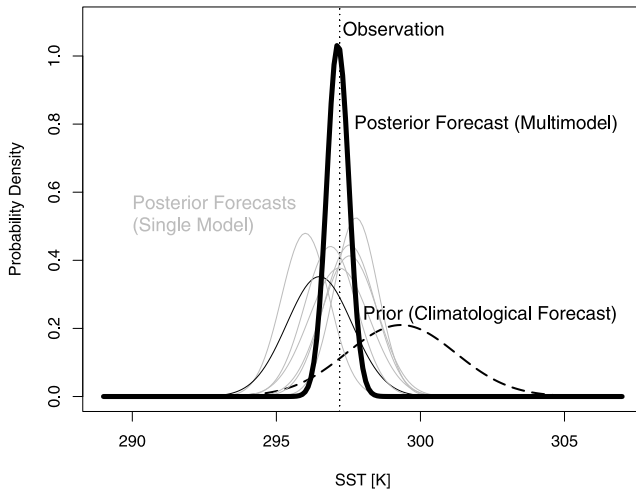
Figure 3. Prior distribution (dashed black), posterior distributions updated with one of the seven DEMETER climate models forecast (solid gray and thin black), and posterior distribution updated with all seven models forecast (solid black) for the forecast of December 1998 monthly mean SST over selected grid (2.5 × 2.5 centered at 0 and 130°W). The thin black line is the posterior distribution (single model) shown in Figure 1.

[17] Figure 3 shows the posterior distribution from the multimodel forecast as well as the posterior distributions from each single model forecast. In this example, it is evident that the mean of the posterior distribution merged from multimodel forecast is closer to the observation than the mean of any of the posterior distributions updated with forecast from only a single model. From equation (10), it can be shown that the variance of the multimodel posterior distribution is always the smallest in all forecasts, as shown in Figure 3. Not surprisingly, when useful information from different model forecasts is gathered, the uncertainty of the forecast should decrease gradually.

[18] In reality, the climate model errors are not completely independent; hence the number of independent models m will be smaller than the actual number of models. Ignoring the correlation among model errors will underestimate the uncertainties, and the extent is dependent on the correlation. To handle cross-model error correlations, a principle component analysis (PCA) can be done through an eigen decomposition of the cross-model error covariance matrix. This would create an orthogonal basis, i.e., a set of linear combinations of regressed model forecasts, where each combination is uncorrelated (independent given the normality assumption) to others. Then the problem reduces to the case of independent errors, and the solution can be easily derived. For the purpose of simplicity, we keep the assumption on model independency in this study.

2.6. Handling of Variables With Nonnormal Distribution

[19] In our proposed Bayesian merging method, both the prior and the likelihood function are assumed to follow normal distributions. Although this makes the mathematical derivation straightforward and produces a normal posterior distribution, the assumption may not be satisfied for many

seasonal prediction variables. This assumption appears valid for monthly mean SST in the example above, even with a relatively small sample size. However, the assumption becomes less suitable for seasonal forecast of precipitation. Daily precipitation over a small region (for example, a 2.5 × 2.5° grid in the midlatitudes) tends not to follow a normal distribution nor does monthly precipitation because the daily distribution tends to have a large mass at zero precipitation, and both daily and monthly distributions are positively skewed. Although the general Bayesian concept for merging information, as expressed by equation (1), still holds, the mathematical derivation of $p(\theta|y)$ becomes more difficult. One potential distributional approach for solving this problem is to base them on a Gamma probability distribution, which has been used in Bayesian analyses [*Wood and Rodriguez-Iturbe*, 1975].

[20] A more general approach for dealing with nonnormally distributed variables is using the method of equal-quantile (cumulative probability) transfer to convert a nonnormal distribution to a normal distribution and vice versa. This equal-quantile approach has been implemented in the work of *Wood et al.* [2002] for a bias correction scheme for seasonal forecasts and is illustrated in Figure 4. The thick black lines are the climatological distributions [PDFs in the upper panels and cumulative density functions (CDFs) in the bottom panels] labeled as "unconditional." The thick black lines in the left panels give the climatological distribution for the variable of interest, for example, the observed monthly precipitation of May over a region within the Ohio River basin. The thick black lines in the right panels are for standard normal distributions. The dashed lines show how variables from the nonnormal distributions are transferred to random variables distributed by a standard normal distribution and vice versa using the equal-quantile principle. Once the climatological distribution of the forecast variable is determined, such a transfer between it and the standard normal is uniquely defined.

[21] The algebraic form of this transfer works as follows. Given an arbitrary nonnormal variable x and its sample $\{x_i\}_{i=1,\ldots,N}$, its CDF or empirical CDF $F_X(\cdot)$ can be obtained by distribution fitting or ranking. Let $F_Z(\cdot)$ be the standard normal CDF, then according to the equal-quantile principle $F_Z(z_i) = F_X(x_i)$, the transfer to the standard normal space is $z_i = F_Z^{-1}(F_X(x_i))$. The Bayesian merging is performed using the standard normal sample $\{z_i\}_{i=1,\ldots,N}$ to obtain the posterior sample $\{z'_i\}_{i=1,\ldots,N}$ (z' is also normal). The inverse transfer is simply $x'_i = F_X^{-1}(F_Z(z'_i))$. If necessary, the distribution of the posterior can be fitted from the sample $\{x'_i\}_{i=1,\ldots,N}$.

[22] This transfer allows us to convert nonnormal variables to normal variates and perform the Bayesian merging on normal variates to obtain the desired conditional posterior distribution. The resulting posterior distribution is a normal distribution (shown by the thin black line in the upper right panel in Figure 4) and is used to estimate the posterior distribution in the climate forecast variable space. This procedure provides the needed information to estimate the posterior distribution and corresponding cumulative in the transformed normal space, and uses the information to estimate the corresponding values in the climate variable space. However, we recognize that the algebraic form of the posterior cannot be predetermined in many cases but that regularly applied techniques for fitting distributions to data
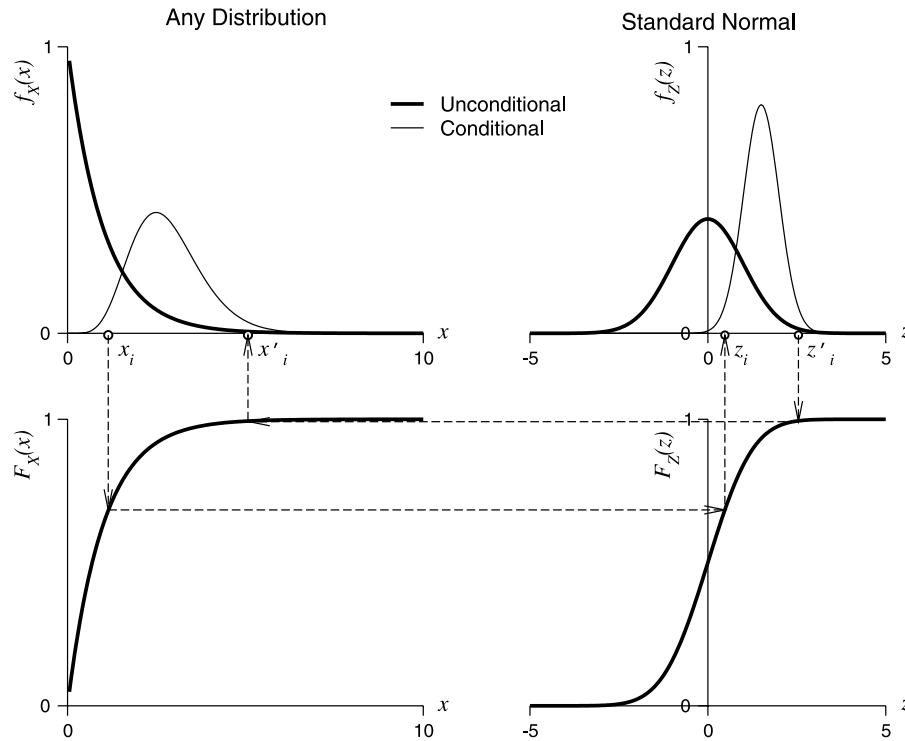
Figure 4. Transfer of a nonnormal distribution to and back from a standard normal distribution. The thick black lines in the left panels are for the nonnormal distribution (climatology) of the variable of interest (PDF on the top and CDF at the bottom), and the thick black lines in the right are for the standard normal. Dashed lines running across show how data values can be transferred back and forth using the equal-quantile principle. Thin black lines (labeled with "Conditional") represent the resulted posterior distribution in the Bayesian merging and how they are converted back to the variable's original climatology space.

can be used. Section 4 will present the application of this transfer method in Bayesian merging of precipitation forecasts and the improvements so obtained.

3. Application of the Bayesian Merging Method on SST Forecast Over Equatorial Pacific Region

[23] The monthly mean SST forecast over a $2.5 \times 2.5°$ box within the Niño-3.4 region for December 1998 is used as an example of the Bayesian merging method developed earlier. In this section, the Bayesian merging method is applied to the SST forecast for all months over the equatorial Pacific region. The raw climate model forecasts are taken from the outputs from the EU DEMETER project [*Palmer et al.*, 2004]. Each year from 1958 to 2001, four 6-month forecasts were made with multiple climate models, starting from February, May, August, and November, respectively. Only the forecasts from 1982 to 1999 are used in this study as all the seven models and SST observations are available during this period. Following the cross-validation principle, the multimodel posterior forecast is computed for each year using parameters estimated from other years. The climatological forecast is also estimated following the same principle. The skills of all forecasts (climatological forecast, climate model raw forecasts, and the multimodel posterior forecasts) are evaluated in two ways. The expected value of the forecast is used as a single-valued deterministic forecast and is evaluated using the root mean square error (RMSE).

The forecast distribution (or more precisely the samples from the forecast distribution) is used to determine forecast probabilities and is evaluated using the ranked probability score (RPS) [*Wilks*, 2006]. Each grid in the region is treated independently; the spatial distribution and the temporal change of the evaluation metrics (RMSE and RPS) help to illustrate the systematic improvement of forecast skill in all locations, seasons, and lead times.

[24] Figure 5 shows the RMSE calculated from all August forecasts. In this case, the expected values of each forecast distribution are used as a single-valued deterministic forecast, although we do not explicitly call it a mean forecast. In Figure 5, the nine panels show the RMSE of the climatological forecast (upper left panel), the raw forecasts from seven models, and the multimodel posterior forecast (lower right panel). The x axis of each panel is longitude, from 175°E to 82.5°W, and the y axis is time, spanning 6 months from August to the next January (lead times $0-5$ months). The same plots for the other three forecast periods show very similar patterns; hence they are not presented here. The following features are evident in these plots:

[25] 1. The RMSE of deterministic climatological forecast varies spatially and seasonally. Because the deterministic climatological forecast predicts the climatological mean, the forecast for a given month does not depend on the lead time, and the RMSE of such a forecast is in fact the standard deviation of the underlying climatological distribution. Over the study region, the largest variation in SST shows up
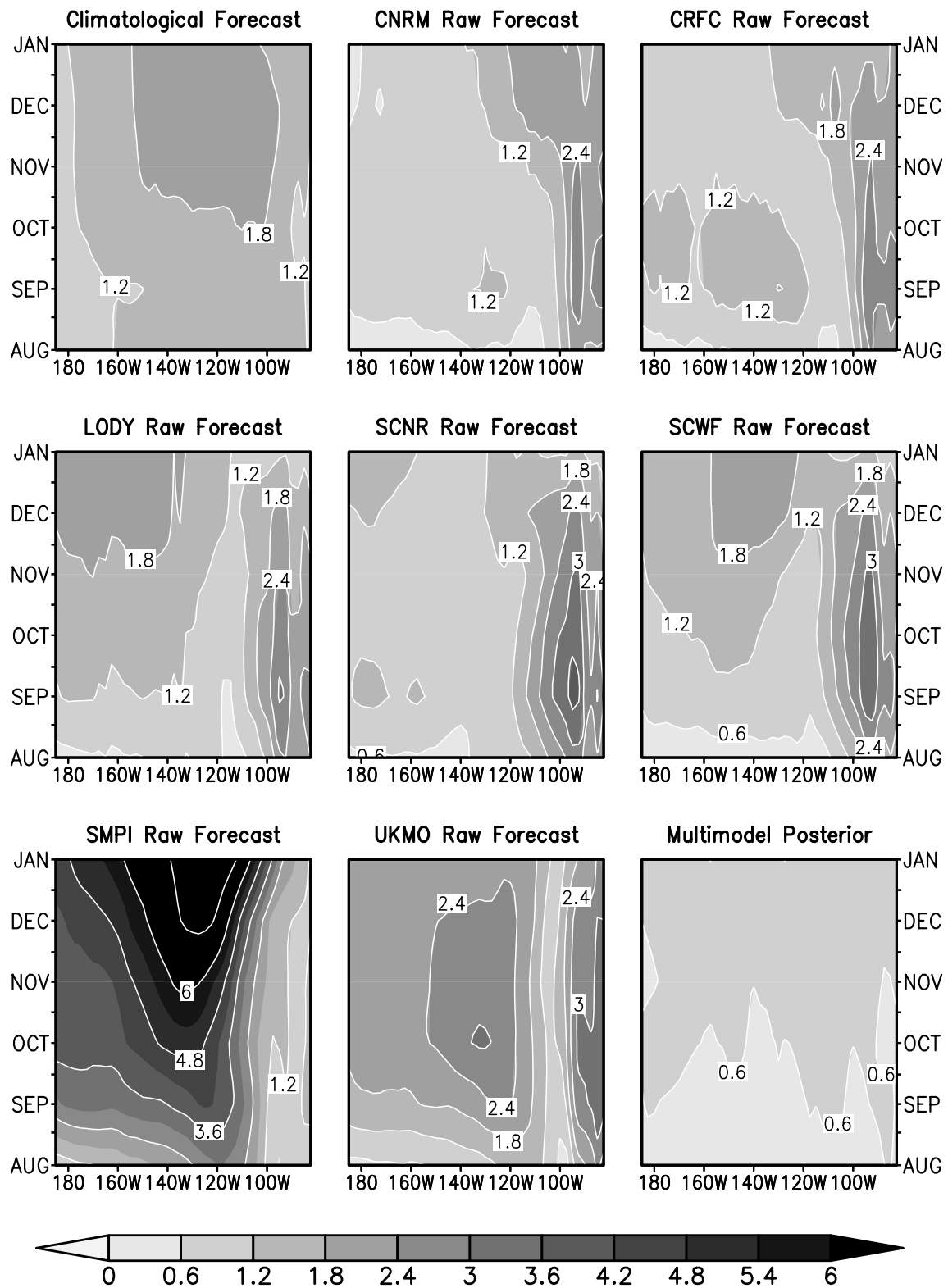
RMSE of AUG SST Forecast

Figure 5. Variation of RMSE with lead time and location from difference forecasts: climatological forecast, raw forecast from seven DEMETER climate models, and the posterior forecast using the seven model forecast and the climatological forecast. This is for all the forecasts starting in August.

during the wintertime over the central part of the region, as a result of the El Niño and Southern Oscillation. Obviously the climatological forecast is incapable of predicting any abnormal events.

[26] 2. The climate model raw forecast RMSE generally grows with lead time over the entire domain. In the DEMETER project, the models are fully coupled atmospheric-ocean models. In the coupled mode, drift in the mean climate states is difficult to avoid, and the general increase in forecast RMSE with lead time partly reflects this drift. The rate of error growth varies with location and model. For example, the error grows faster in the middle and western part of the domain in forecasts from Max-Planck Institute (SMPI) and United Kingdom Meteorological Office models, while the error grows faster in the eastern part of the domain in the other five models. This might be due to the coupling strength between the oceanic and atmospheric models and complex feedbacks within the system.

[27] 3. Among all forecasts, the multimodel posterior forecast shows the smallest RMSE with values around 0.6–1.2 K, while the single model forecast RMSEs are always larger than 0.6 K and up to 7 K for the SMPI model forecast. This is a significant improvement in overall forecast skill. More importantly, there is no obvious error growth with lead time across the entire domain for the multimodel posterior forecast, which suggests that the improvement is systematic regardless of lead time.

[28] 4. As a result of feature 3, the error patterns in the climatological forecast and the raw model forecast are not inherited in the multimodel posterior forecast. This is particularly important to the El Niño forecast as seen in the August forecast. While the climatological forecast shows the largest error around December in the middle part of the domain, forecast errors from climate models show significantly different patterns. For example, most raw model forecasts show the largest error in the eastern part of the domain during September, while one model (SMPI) shows the largest error in the central part of the domain in January with a value of 7 K. However, in the multimodel posterior forecast, all aforementioned error patterns disappear, and errors have been reduced to less than 1 K, almost constant across the domain for all periods.

[29] Figure 6 shows the evaluation of these forecasts in a probabilistic manner using the Ranked Probability Skill Score (RPSS). The ranked probability score is essentially an extension of the Brier score applied to evaluating many events [*Wilks*, 2006] and thus considers not only the location of the mean probabilistic forecast but also the spread of the probabilistic forecasts. A perfect forecast would assign all the probability to the single category corresponding to the event that subsequently occurs, so RPS = 0. The RPSS for each forecast model can be computed with respect to a reference forecast. A value of RPSS of 1 indicates perfect multicategory probabilistic forecasts, while a value of 0 or less means the forecast is not superior to the reference forecast. In Figure 6, the RPS from the multimodel posterior forecast is used as the reference to calculate RPSS for the climatological forecast and individual climate model forecasts. Large negative values of RPSS shown in the plots indicate that the reference forecast, i.e., the multimodel posterior forecast,

is significantly better than the prior (climatological forecast) and any individual model's raw forecast.

[30] Different climate models have individual strengths in predicting SST in different regions at different times, as presented by the RMSE patterns and RPSS pattern from their forecasts. A good postprocessing technique should pick up the strengths and combine them together to produce the best forecast. Figure 7 shows the average contribution of each model to the multimodel posterior forecast. The contribution is represented as $\beta_i^2/(\varphi_{y_i} + \varphi_{\varepsilon_i})$ as indicated by equation (10). Figure 7 suggests that climate models are more skillful at short lead times; hence they contribute more toward the posterior forecast. With the increase in lead time, the contributions from all models decrease and the contribution from the prior (climatology) dominates. This is exactly what we want to achieve with the Bayesian merging method, and Figure 7 clearly illustrates it. Figures 5 and 6 clearly demonstrate that the Bayesian merging method is capable of producing significantly improved SST seasonal forecasts by statistically combining climate model raw forecasts with observed climatology. It removes the biases in individual model forecast that may vary spatially across the study region and varies temporally with lead time and season. The posterior of the multimodel Bayesian merging approach, as a multimodel forecast model, has the probability focused on the right place (subsequent observations) so that the mean forecast error is smaller and the confidence over the climatological and individual climate model forecasts is higher.

4. Application of the Bayesian Merging Method to Precipitation Forecasting Over Ohio River Basin

[31] The Bayesian merging method is also applied to monthly precipitation forecasts over the Ohio River basin. Because of the skewed distribution of monthly precipitation, the equal-quantile transfer scheme described in section 2.6 is applied. The model setting is exactly the same as in the SST forecast above, but the target is the May to October monthly precipitation. As an example, only one $2.5 \times 2.5°$ grid is considered here, which is centered at 40°N and 82.5°W in the eastern portion of the Ohio River basin. The precipitation forecasts from the DEMETER models are used to merge with observed climatology, which is derived from the high-resolution gridded precipitation data set produced by the Climate Research Unit at the University of East Anglia [*Mitchell et al.*, 2003] and is regridded to match the DEMETER model grid. All nonnormal distributions are transferred to the standard normal distributional space before applying the Bayesian merging method. After the Bayesian merging, the posterior distribution of the precipitation forecast is transferred back to the precipitation space. These forecasts are cross validated in exactly the same way as in the SST forecast.

[32] Figure 8 presents the time series of May monthly precipitation for the 19-year period from the different forecasts: the climatological forecast, seven climate model forecasts, and the multimodel posterior forecast. May is the first month of the 6-month forecast initialized from 1 May. The expected value of each forecast distribution is used in the deterministic forecast verification. For the 19 Mays, the RMSEs of the seven climate model forecasts range between
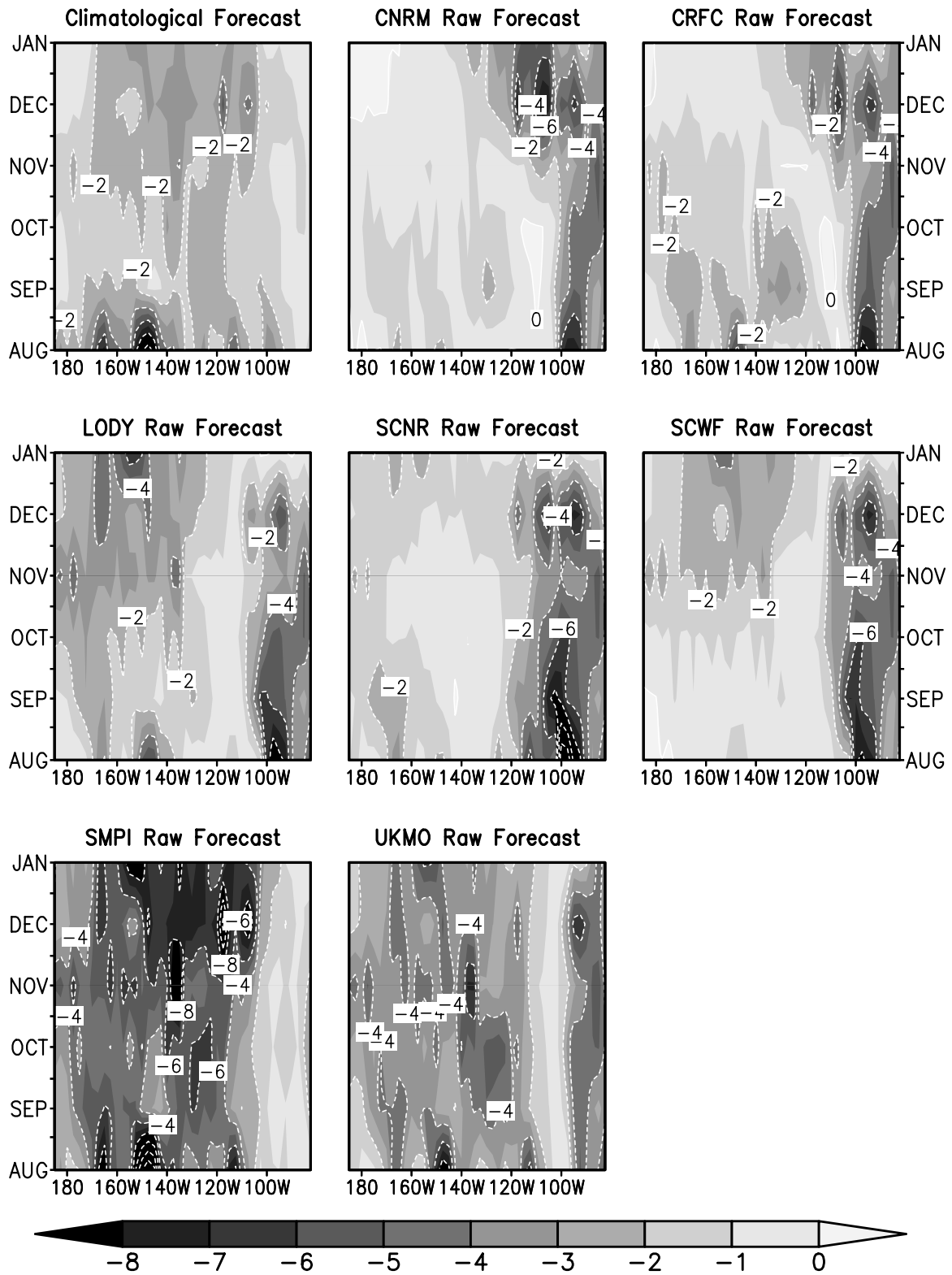
Figure 6. Variation of average RPSS with lead time and location from climatological forecast and raw forecast by seven DEMETER climate models. The reference RPS is the multimodel posterior forecast from the seven climate model forecast and the climatology. This is for all the forecasts starting in August.
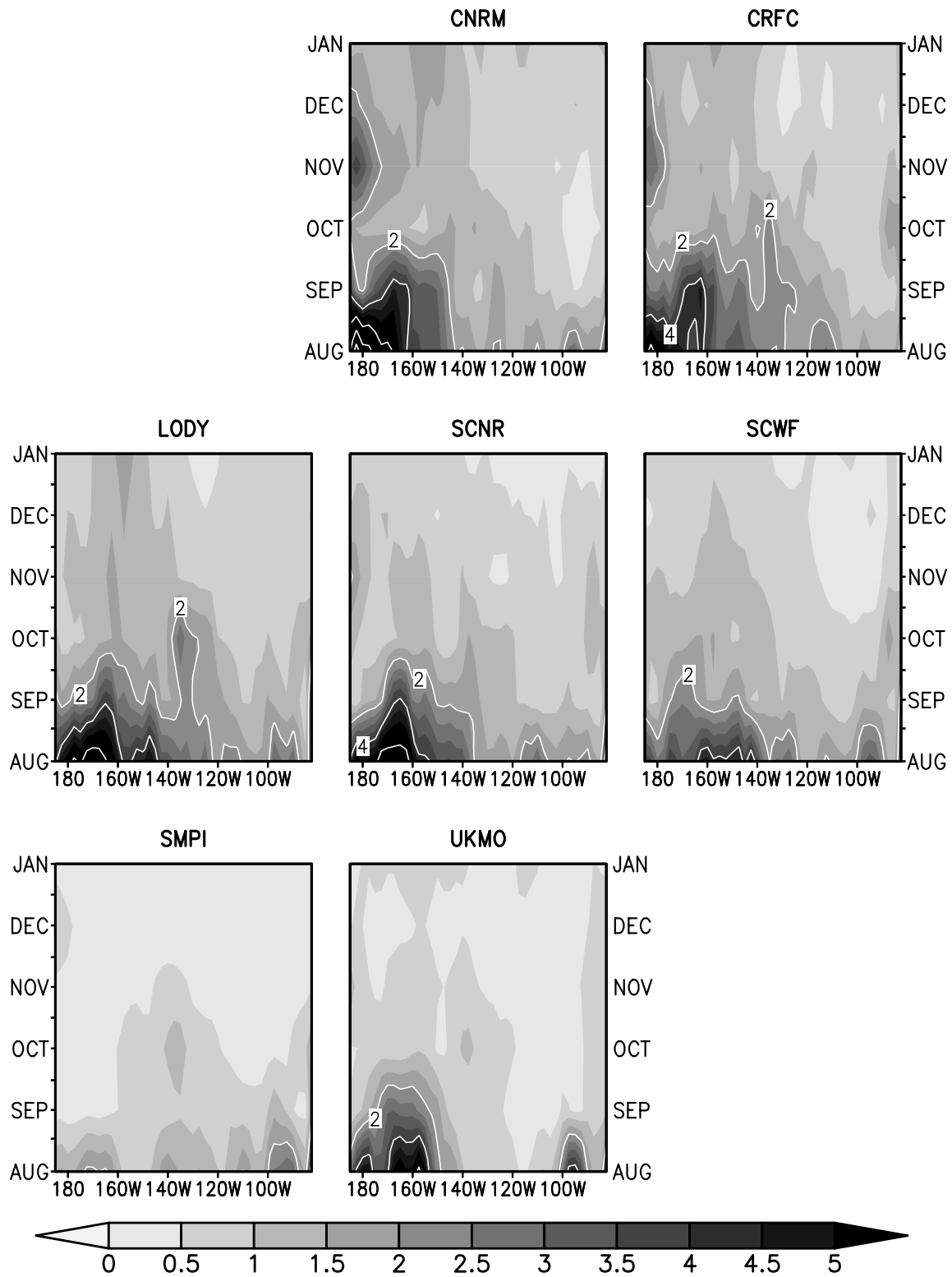
Figure 7. Variation of average contribution of each model to the multimodel posterior forecast with lead time and location. This is for all the forecasts starting in August.
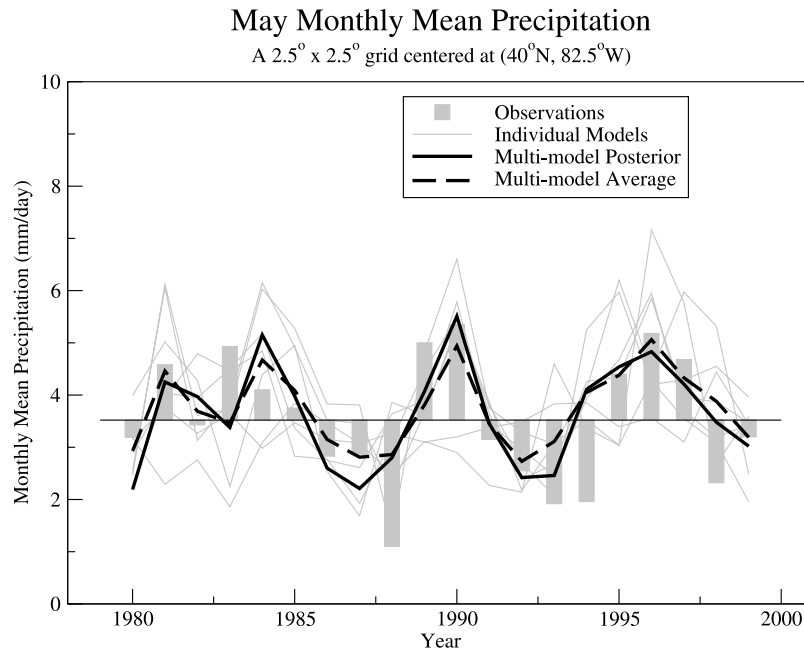
May Monthly Mean Precipitation
A 2.5° x 2.5° grid centered at (40°N, 82.5°W)

Figure 8. Multimodel posterior forecast (solid black line) and observed (gray bars) precipitation for all the May at the selected grid with the Ohio River basin, plotted relative to the climatological mean as indicated by the horizontal line. The observations are obtained from the CRU data set and regridded to match the climate model grid. The ensemble mean of each climate model forecast is plotted as solid gray lines. The dashed black line is the average of all model means. The forecasts are the first month of each 6-month forecast starting from 1 May.

0.94 and 1.49 mm/day, while the RMSE of the multimodel posterior forecast model is about 0.89 mm/day. Although the merged forecast is not always the best, the forecast error has been reduced and skill has been obtained. When evaluated as a probabilistic forecast, the multimodel posterior also shows higher forecast skills represented by a smaller RPS. The average RPS of the climatological forecast and the multimodel posterior forecast is 0.88 and 0.64, respectively. The average RPS of the climate model raw forecasts range between 0.66 and 1.15. For comparison, we also show the simple average of the multiple model forecasts in Figure 8, and it is very close to the multimodel posterior forecast. The simple averaging suggests equal weights for the seven models, so the multimodel posterior forecast should be no worse than the simple average assuming that a model receiving a higher weight reflects higher model skill. In this case, the simple average is the multimodel posterior mean with an RMSE of 0.9 mm/day.

[33] Figure 8 shows the 1-month ahead forecast. The mean of the multimodel posterior distribution differs from the climatological forecast, and the multimodel posterior distribution is also different from the climatological distribution, which have different values of RPS. When using the climatological forecast as the reference forecast, the RPSS of the multimodel posterior forecast is 0.26. However, when the lead time increases, the multimodel posterior distribution is not significantly different from the climatological distribution, indicating that the dynamical climate model forecasts have little skill in predicting precipitation over this region beyond 1 month. *Luo and Wood* [2006] showed that the NCEP Climate Forecast System has no potential predictability of monthly precipitation at this spatial scale with

lead times longer than 1 month, and the results above show this is true for the DEMETER project models. Naturally, this lack of potential predictability results in unskillful forecasts. In our multimodel approach, the lack of skill is automatically taken into account by the Bayesian merging method in the likelihood function (smaller β and large φ_ε), so the contribution from the (unskillful) climate models to the posterior distribution is tiny during the Bayesian update. The posterior distribution therefore is not significantly different from the prior distribution, i.e., the climatological distribution.

5. Discussion and Conclusions

[34] The value of seasonal climate forecasts has been increasingly recognized over the last decade or so, and many statistical models along with dynamical models have been developed and used in seasonal forecasting. Understanding the predictive skill of the individual models, classes of models (statistical and dynamical), and the best method for applying their forecasts has been a challenge to the forecast community. In this paper, we develop a multimodel forecast system based on combining competing forecasts using a Bayesian merging approach. It was shown that such postprocessing of seasonal climate forecasts to produce a merged forecast achieves better skill than any individual forecast model. The proposed approach develops the likelihood function using linear regressions of observations and mean ensemble hindcasts (historical forecasts). The regressions effectively remove biases in the model forecasts. By quantifying the model skills with regression error variances as well as the ensemble spread and combining

this with the climatological distribution of observations that offers a prior forecast, the Bayesian merging approach extracts useful information from each source to produce the minimum variance posterior forecast. The method has been tested using forecasts from seven climate forecast models from the EU DEMETER project, first, for SST forecasts over the Equatorial Pacific region and, second, for monthly precipitation for a grid cell within the Ohio River basin. Results show that the posterior forecast using the multimodel forecast has the lowest RMSE and lowest RPS. One interesting side bar from the Bayesian approach is the observation that forecasts based on the prior climatological distribution was more skillful in the eastern portion of the domain at longer lead times than for individual model forecasts, which is not true for the multimodel Bayesian forecasts.

[35] It is necessary to understand that the multimodel posterior forecast provides the "expected" best (i.e., smallest RMSE) forecast and will therefore perform the best over many forecast cycles. For specific forecasts, one dynamic model might happen to give a perfect forecast, while its past performance might provide little evidence for us to trust its current forecast. Therefore in evaluating the Bayesian framework, both the past performance and the current forecast are important. The sample size in estimating the coefficients in the linear model, i.e., the number of forecasts used in the regression, also has an impact on the Bayesian merging method in that a large sample size (long set of hindcasts) will provide a more accurate estimate of the likelihood function. Therefore a long-term hindcast data set with the same dynamic climate model is necessary in developing better forecast systems.

[36] We also propose a simple method to extend the application of the Bayesian merging method to seasonal hydrological predictions, including forecasting of monthly precipitations in the midlatitudes. In applying the Bayesian approach to seasonal hydrological predictions, specifically forecasting of monthly precipitations in the midlatitudes, the challenges include the skewed distributions of the forecast variable (both in climatological distribution and forecast ensemble), which was overcome with the proposed equal-quantile transfer method to convert all skewed distributions to normal (nonskewed) distributions. The example shown in this study illustrates the potential of using the Bayesian merging method in seasonal hydrological predictions. The posterior forecast of monthly precipitation for the Ohio River basin grid from multiple models is not significantly better than the simple average of multimodel forecast, which results from uniform predictive skill among the models. The posterior is not significantly different from the prior climatological distribution, showing the lack of predictive skill by the dynamical models for midlatitude monthly precipitation at the scales studied.

[37] Another attractive feature of the Bayesian merging method is that the spatial and temporal scale of the variable θ and the model forecast y are not required to be the same. Thus θ may represent precipitation over a small basin, and y may represent a forecast made at the seasonal climate model grid scale overlying the basin. Thus our Bayesian approach offers an effective way to statistically downscale information from large scales, which normally comes from climate model forecasts, to smaller scales that are suitable for hydrological applications. L. Luo and E.F. Wood (Seasonal

hydrological prediction with the VIC hydrologic model for the Ohio River basin, in preparation for submission to Journal of Hydrometeorology, 2007) implemented this method in their seasonal hydrological ensemble prediction system and showed skillful seasonal forecasts of soil moisture and streamflow over the Ohio River basin.

[38] We expect that the proposed Bayesian merging method can be further improved when the dependency among multiple climate models are handled more carefully. For simplicity, we assume that the errors from the seven DEMETER models are independent, but indeed they are not. A principle component analysis (not shown here) on the monthly precipitation forecast errors from these models shows that the first three eigenvectors can explain about 58, 17, and 11% of the total variance; therefore the effective number of independent models is less than seven. Not considering correlations among model errors in the multimodel Bayesian merging results in overweighting the climate model forecasts and underweighting the climatology (prior) distribution. When the dynamical climate models are not very skillful, this leads to poorer forecast performance. Approaches for handling such correlated model forecasts are available; for example, approaches using multivariate normals within a Bayesian framework can be found in the work of *Zellner* [1971] and *Stephenson et al.* [2005].

[39] The proposed Bayesian merging method appears to have great potential in postprocessing multimodel ensemble forecasts, as demonstrated using both seasonal tropical SST forecasts and seasonal precipitation forecasts. The proposed approach extends the super-ensemble approach of *Krishnamurti et al.* [2000] by considering model weights dependent on model skill and offers a consistent approach when the model forecasts are combined with climatology. A straightforward extension to the approach is the merging of dynamical and statistical climate model forecasts, with forecasts based on climatology, and handling the various spatial forecast scales offered by the suite of models being merged. Such a multimodel Bayesian forecast would provide improved deterministic and probabilistic forecasts for users and decision makers from what already exists.

References

Barnston, A. G., A. Leetmaa, V. E. Kousky, R. E. Livezey, E. A. O'Lenic, H. Van den Dool, A. J. Wagner, and D. A. Unger (1999), NCEP forecasts of the El Nino of 1997–98 and its US impacts, *Bull. Am. Meteorol. Soc.*, *80*, 1829–1852.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak (2003), Multi-model ensembling in seasonal climate forecasting at IRI, *Bull. Am. Meteorol. Soc.*, *84*, 1783–1796.

Berliner, L. M., C. K. Wikle, and N. Cressie (2000), Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling, *J. Clim.*, *13*, 3953–3968.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson (2003), Skill of coupled model seasonal forecasts: A Bayesian assessment of ECMWF ENSO forecast. ECMWF Technical Memorandum No. 426, 16pp., ECMWF, Reading, UK.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson (2004), Forecast calibration and combination: A simple Bayesian approach for ENSO, *J. Clim.*, *17*, 1504–1516.

Everinghama, Y. L., R. C. Muchow, R. C. Stone, N. G. Inman-Bamber, A. Singels, and C. N. Bezuidenhout (2002), Enhanced risk management and decision-making capability across the sugarcane industry

value chain based on seasonal climate forecasts, *Agric. Syst.*, *74*, 459–477.

Hoffman, R. N., and E. Kalnay (1983), Lagged average forecasting, an alternative to Monte Carlo forecasting, *Tellus*, *35a*, 100–118.

Jones, J. W., J. W. Hansen, F. S. Royce, and C. D. Messina (2000), Potential benefits of climate forecasting to agriculture, *Agric. Ecosyst. Environ.*, *82*, 169–184.

Kanamitsu, M., A. Kumar, H. M. H. Juang, J. K. Schemm, W. Q. Wang, F. L. Yang, S. Y. Hong, P. T. Peng, W. Chen, S. Moorthi, and M. Ji (2002), NCEP dynamical seasonal forecast system 2000, *Bull. Am. Meteorol. Soc.*, *83*, 1019–1037.

Koster, R. D., M. J. Suarez, and M. Heiser (2000), Variance and predictability of precipitation at seasonal-to-interannual timescales, *J. Hydrometeorol.*, *1*, 26–46.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. Larow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000), Multi-model ensemble forecasts for weather and seasonal climate, *J. Clim.*, *13*, 4196–4216.

Krishnamurti, T. N., et al. (2001), Real-time multianalysis-multi-model superensemble forecasts of precipitation using TRMM and SSM/I products, *Mon. Weather Rev.*, *129*, 2861–2883.

Kumar, T. S. V. V., T. N. Krishnamurti, M. Fiorino, and M. Nagata (2003), Multi-model superensemble forecasting of tropical cyclones in the Pacific, *Mon. Weather Res.*, *131*, 574–583.

Landsea, C. W., and J. A. Knaff (2000), How much skill was there in forecasting the very strong 1997–98 El Nino?, *Bull. Am. Meteorol. Soc.*, *81*, 2017–2119.

Latif, M., M. D. Anderson, T. Barneet, M. Cane, A. Leetmaa, J. O'Brien, A. Rosati, and E. Schneider (1998), A review of the predictability of ENSO, *J. Geophys. Res. [Ocean]*, *103*, 14,375–14,393.

Lee, P. M. (1997), *Bayesian Statistics: An Introduction*, 2nd ed., 344 pp., Edward Arnold, London.

Luo, L., and E.F. Wood (2006), Assessing the idealized predictability of precipitation and temperature in the NCEP Climate Forecast System, *Geophys. Res. Lett.*, *33*, L04708, doi:10.1029/2005GL025292.

Mitchell, T. D., et al. (2003), A comprehensive set of climate scenarios for Europe and the globe, *Tyndall Centre Working Paper*, *55*.

Palmer, T. N. (2002), The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, *Q. J. R. Meteorol. Soc.*, *128*, 747–774.

Palmer, T. N., C. Brankovic, and D. S. Richardson (2000), A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations, *Q. J. R. Meteorol. Soc.*, *126*, 2013–2033.

Palmer, T. N., et al. (2004), Development of a European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, *85*, 853–872.

Rajagopalan, B., U. Lall, and S. E. Zebiak (2002), Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles, *Mon. Weather Rev.*, *130*, 1792–1811.

Reynolds, R. W., and T. M. Smith (1994), Improved global sea surface temperature analyses using optimum interpolation, *J. Clim.*, *7*, 929–948.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard (2004), Improved combination of multiple atmospheric GCM ensembles for seasonal prediction, *Mon. Weather Rev.*, *132*, 2732–2744.

Schneider, J. M., and J. D. Garbrecht (2003), A measure of the usefulness of seasonal precipitation forecasts for agricultural applications, *Trans. ASAE*, *46*, 257–267.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda (2005), Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions, *Tellus A*, *57*, 253–264.

Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, *Bull. Am. Meteorol. Soc.*, *74*, 2317–2330.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*. 2nd ed., 627 pp., Elsevier, New York.

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. V. Kumar (2003), Real-time multi-model superensemble forecasts of Atlantic tropical systems of 1999, *Mon. Weather Rev.*, *131*, 1878–1894.

Wood, E. F., and I. Rodriguez-Iturbe (1975), Bayesian inference and decision making for extreme hydrologic events, *Water Resour. Res.*, *11*, 533–542.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier (2002), Long range experimental hydrologic forecasting for the eastern U.S., *J. Geophys. Res.*, *107*(D20), 4429, doi:10.1029/2001JD000659.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti (2003), Improvement of the multi-model superensemble technique for seasonal forecasts, *J. Clim.*, *16*, 3834–4830.

Zellner, A. (1971), *Introduction to Bayesian Inference in Econometrics*, 431 pp., John Wiley, Hoboken, N. J.

————————————

L. Luo, Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ 08544, USA. (lluo@princeton.edu)

M. Pan and E. F. Wood, Environmental Engineering and Water Resource, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544, USA.